

Data Documentation Initiative

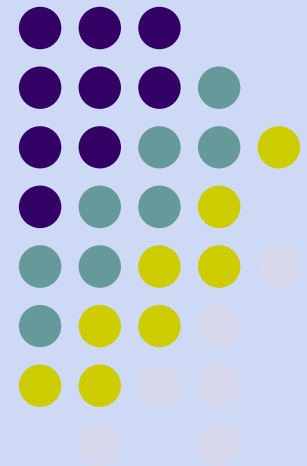
DDI

Ann Green

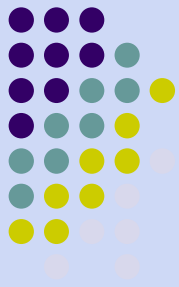
Digital Life Cycle Research & Consulting

green.ann@gmail.com

dlifecycle.net



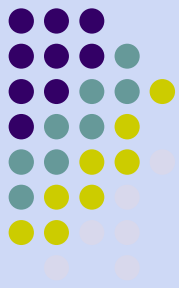
Presented at the DataShare project meeting
University of Edinburgh, Feb 5-6, 2008



Evolution of the versions of DDI

- DDI 1: microdata surveys
- DDI 2: added aggregate tabular data
- DDI 3: modular, life cycle model, complex data files, comparative data files

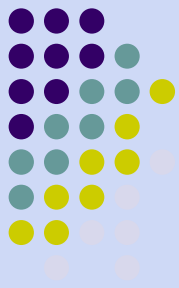




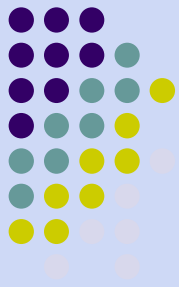
Structure of DDI2

- **Document description:** information about the DDI document and how it was created with bibliographic citation
- **Study description:** information about the context of the data production and distribution (creators, methodology, abstract, keywords, etc.)
- **Data files description:** information about the data file or files (format, size, number of cases, etc.)
- **Variable description:** information about the data items or rows and columns in a tabular data file/s
- **Other study materials:** inline reference materials or references to external reference materials (coding schemes, thesauri, citations to publications, etc.)

DDI 2 and DDI 3 comparison



Version 2	Version 3.0
Inadequate representation of complex / hierarchical data	Detailed documentation for complex / hierarchical data
No instrument coverage.	Full description of instrument as a separate entity.
Question text appears only as part of variable description.	Compatible with Computer Assisted Interviewing software.
No documentation for question flow / conditions. Initially designed for microdata only	Documents specific use of questions: flow, conditions, loops. Adds support for tabular, spreadsheet-type, representation of aggregate data
Aggregate data section added in V 2.1 to support limited representation (Census-type data, delimited files) No data transport function	Aggregate data transport option: cell content may be included inline with the data item description In-line inclusion enabled for both aggregate data and microdata
No Longitudinal / Time Series / Cross-national Data Comparability	Grouping structure documents studies related on one or several dimensions (time, geography, language, etc.) as well as their comparability
Limited Multilingual Support	Support for multiple language use and translations
Single File, Hierarchical design	Modular design: Facilitates reuse; Facilitates versioning and maintenance; Supports life cycle model; Allows flexibility in organizing the DDI Instance; Supports grouping and comparing studies; Supports creation of metadata registries



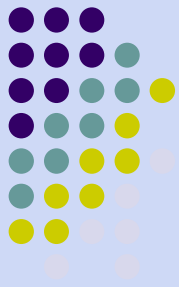
DDI 3 timetable for publication

Version 3.0:

- February 15 - April 14, 2007: Public Review Period
- April 15 - May 1, 2007: Adjusted Schemas based on Public Review
- May 2007: Voted to Publish Version 3.0 as Candidate Draft
- July 2007: Candidate Draft Released
- December 2007: Full Proof of Concept Presented
- April 2008: Official Publication of Version 3.0; Updated examples and use cases; High level documentation updated

<http://www.ddialliance.org/ddi3/index.html>

Version 3.0 is Life Cycle oriented and modular



Designed to cover all stages in the life cycle of a data collection:

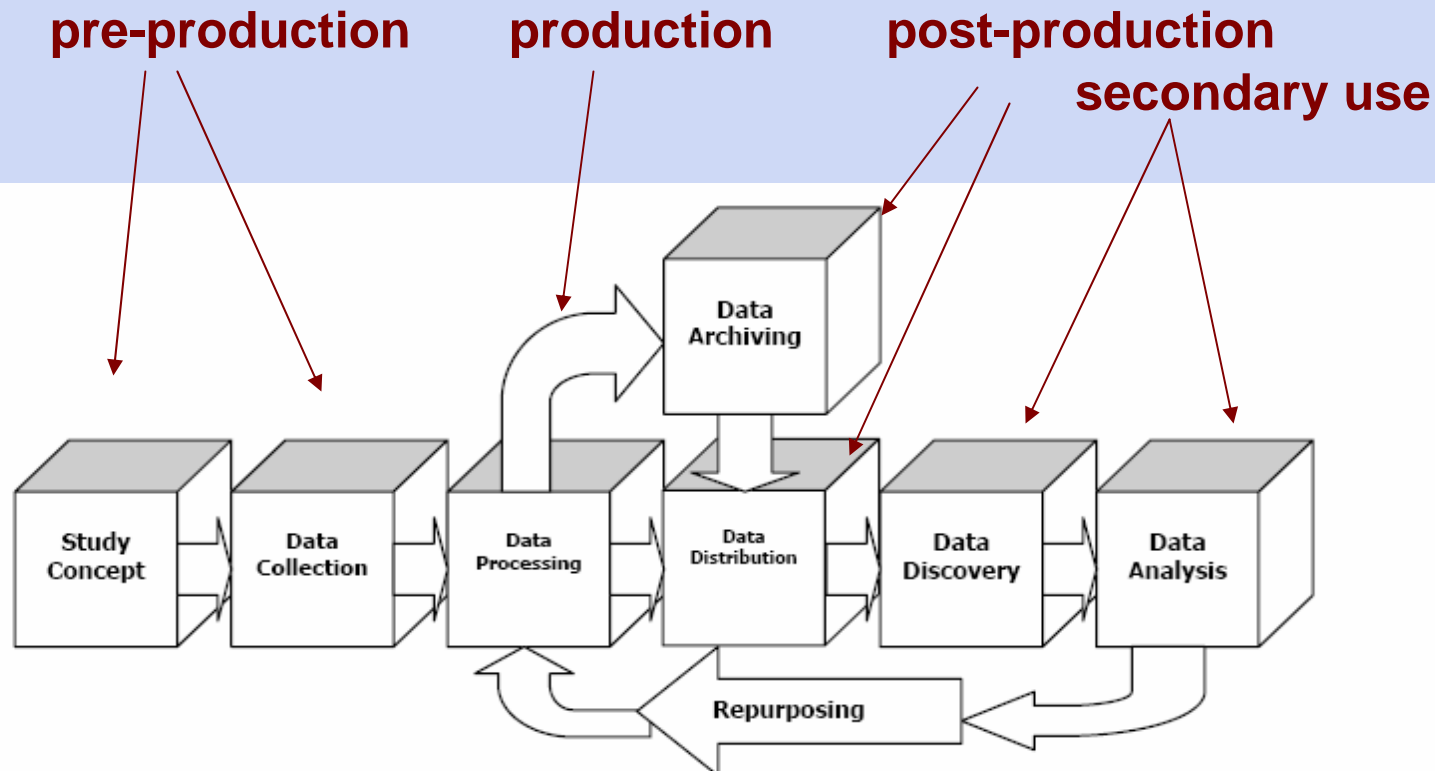
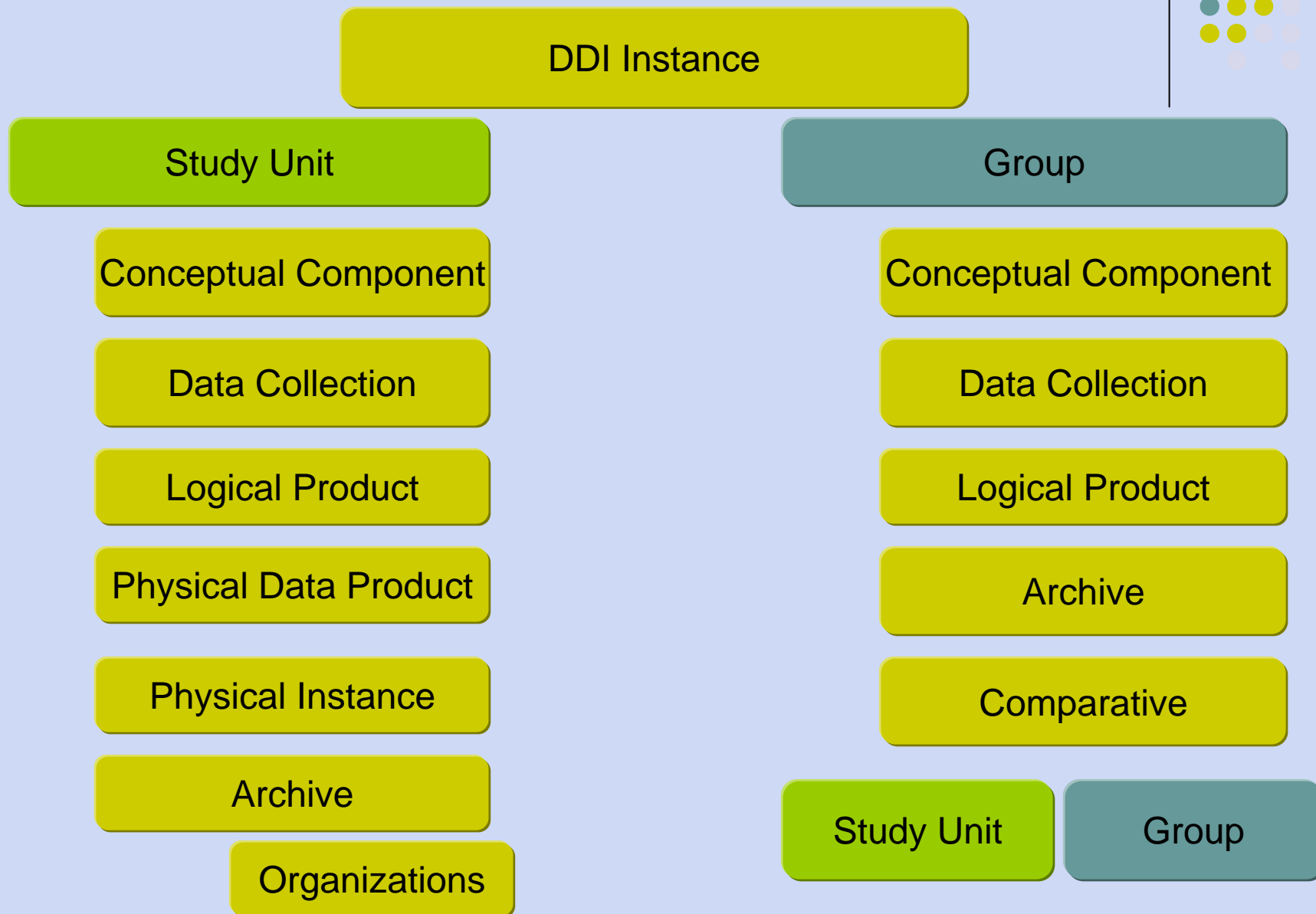
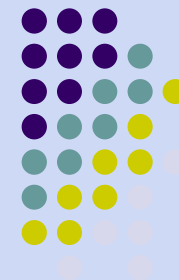
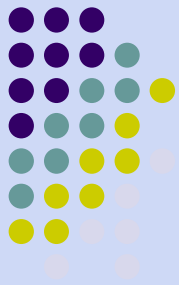


Figure: Combined Life Cycle Model

DDI Version 3.0 Modules

-- Structural Overview --

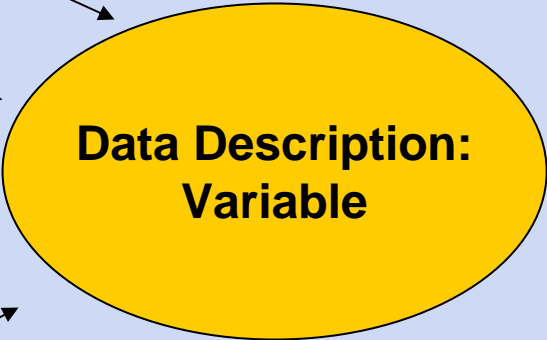


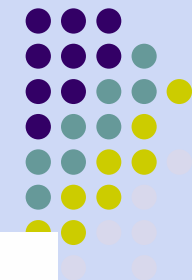


Version 2.1 vs. Version 3.0

Example: A survey variable in Version 2.1

```
V043015 A7a. Attention to national (network) news
-----
=====
PRE-ELECTION SURVEY:
IF R WATCHED NATIONAL NETWORK TV NEWS IN PAST WEEK:
QUESTION:
-----
Please look at page 1 of the booklet.
How much attention do you pay to news on NATIONAL news
shows about the campaign for President -- a GREAT DEAL,
QUITE A BIT, SOME, VERY LITTLE, or NONE?
VALID CODES:
-----
1. A great deal
2. Quite a bit
3. Some
4. Very little
5. None
MISSING CODES:
-----
8. Don't know
9. Refused
INAP. 0,8,9 in A7
TYPE:
-----
Numeric Dec 0
. 270
1 198
2 318
3 304
4 110
5 12
```





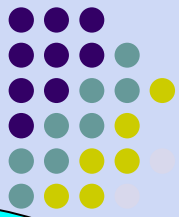
Version 2.1 vs. Version 3.0

Example: A survey variable in Version 2.1

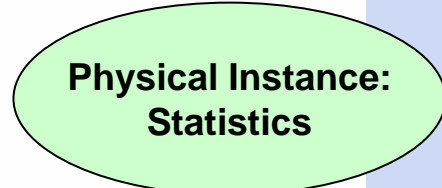
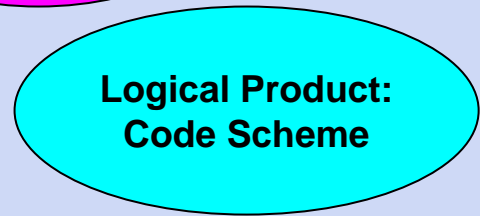
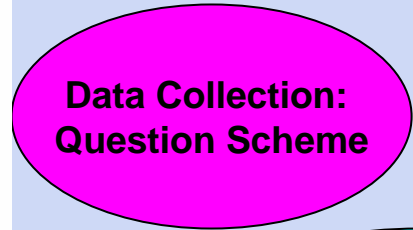
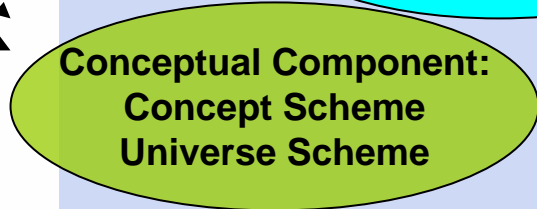
```
var location/ < > name="V043015"
labl Attention to national (network) news /labl
qstn qstnLit Please look at page 1 of the booklet. How much attention do you pay to news on NATIONAL news shows about
the campaign for President -- a GREAT DEAL, QUITE A BIT, SOME, VERY LITTLE, or NONE? /qstnLit /qstn
valrng range/ < > /valrng
universe Pre-election survey: respondents who watched national network TV news past week (1-7 in V043014) /universe
sumStat 942 /sumStat
catgry labl INAP /labl
catStat 270 /catStat /catgry
catgry catValu 1 /catValu
labl A great deal /labl
catStat 198 /catStat /catgry
catgry catValu 2 /catValu
labl Quite a bit /labl
catStat 318 /catStat /catgry
catgry catValu 3 /catValu
labl Some /labl
catStat 304 /catStat /catgry
catgry catValu 4 /catValu
labl Very little /labl
catStat 110 /catStat /catgry
catgry catValu 5 /catValu
labl None /labl
catStat 12 /catStat /catgry
catgry catValu 8 /catValu
labl Don't know /labl
catStat 0 /catStat /catgry
catgry catValu 9 /catValu
labl Refused /labl
catStat 0 /catStat /catgry
concept Attention to presidential campaign on national TV /concept
varFormat numeric /varFormat /var
```

Version 2.1 vs. Version 3.0

Example: A survey variable in Version 3.0

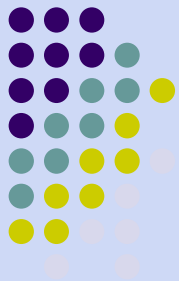


```
V043015 A7a. Attention to national (network) news
-----
=====
PRE-ELECTION SURVEY:
IF R WATCHED NATIONAL NETWORK TV NEWS IN PAST WEEK:
QUESTION:
-----
Please look at page 1 of the booklet.
How much attention do you pay to news on NATIONAL news
shows about the campaign for President -- a GREAT DEAL,
QUITE A BIT, SOME, VERY LITTLE, or NONE?
VALID CODES:
-----
1. A great deal
2. Quite a bit
3. Some
4. Very little
5. None
MISSING CODES:
-----
8. Don't know
9. Refused
INAP. 0,8,9 in A7
TYPE:
-----
Numeric Dec 0
. 270
1 198
2 318
3 304
4 110
5 12
```



DDI 3.0 Markup

Modules used in a full variable description



```

<c:ConceptualComponent> <r:MaintainableID> <r:ID> 4245_ConceptualComponent </r:ID> </r:MaintainableID>
<c:ConceptScheme> <r:MaintainableID> <r:ID> 4245_ConceptScheme </r:ID> </r:MaintainableID>
<c:Concept> <r:VersionableID> <r:ID> Concept_2 </r:ID> </r:VersionableID>
<r:Description> Attention to presidential campaign on national TV </r:Description> </c:Concept> </c:ConceptScheme>
<c:UniverseScheme> <r:MaintainableID> <r:ID> 4245_UniverseScheme </r:ID> </r:MaintainableID>
<c:Universe> <r:MaintainableID> <r:ID> 4245_Universe </r:ID> </r:MaintainableID>
<r:HumanReadable> <xhtml:p> All United States citizens of voting age on or before the 2004 Election Day. </xhtml:p> </r:HumanReadable>
<r:SubUniverse> <r:MaintainableID> <r:ID> SubUniverse1_Preelection </r:ID> </r:MaintainableID>
<r:HumanReadable> <xhtml:p> Respondents who watched national network TV news during the previous week </xhtml:p>
</r:HumanReadable> </r:SubUniverse> </c:Universe> </c:UniverseScheme> </c:ConceptualComponent>
<d:DataCollection> <r:MaintainableID> <r:ID> 4245_DataCollection </r:ID> </r:MaintainableID>
<d:QuestionScheme> <r:MaintainableID> <r:ID> 4245_QuestionScheme </r:ID> </r:MaintainableID>
<d:QuestionItem> <r:IdentifiableID> <r:ID> A7a </r:ID> </r:IdentifiableID>
<d:QuestionText> <d:LiteralText> <d:Text> Please look at page 1 of the booklet. How much attention do you pay to news on
NATIONAL news shows about the campaign for President -- a GREAT DEAL, QUITE A BIT, SOME, VERY LITTLE, or NONE? </d:Text>
</d:LiteralText> </d:QuestionText>
<d:CategoryDomain> <r:CategorySchemeReference> <r:Reference> <r:ID> CategoryScheme_V043015 </r:ID> </r:Reference>
</r:CategorySchemeReference> </d:CategoryDomain> </d:QuestionItem> </d:QuestionScheme> </d:DataCollection>
<l:LogicalProduct> <r:MaintainableID> <r:ID> 4245_LogicalProduct </r:ID> </r:MaintainableID>
<l:CategoryScheme> <r:MaintainableID> <r:ID> 4245_CategoryScheme </r:ID> </r:MaintainableID> </l:CategoryScheme>
<l:CodeScheme> <r:MaintainableID> <r:ID> CodeScheme_V043015 </r:ID> </r:MaintainableID>
<l:CategorySchemeReference> <r:Reference> <r:ID> 4245_CategoryScheme </r:ID> </r:Reference> </l:CategorySchemeReference> </l:CodeScheme>
<l:VariableScheme> <r:MaintainableID> <r:ID> 4245_VariableScheme </r:ID> </r:MaintainableID>
<l:Variable> <r:IdentifiableID> <r:ID> V043015 </r:ID> </r:IdentifiableID>
<r:Label> Attention to national (network) news </r:Label>
<r:UniverseReference> <r:Reference> <r:ID> SubUniverse1_Preelection </r:ID> </r:Reference> </r:UniverseReference>
<l:ConceptReference> <r:Reference> <r:ID> Concept_2 </r:ID> </r:Reference> </l:ConceptReference>
<l:QuestionReference> <r:Reference> <r:ID> A7a </r:ID> </r:Reference> </l:QuestionReference>
<l:Representation> <l:CodeRepresentation> <r:CodeSchemeReference> <r:Reference> <r:ID> CodeScheme_V043015 </r:ID> </r:Reference>
</r:CodeSchemeReference> </l:CodeRepresentation> </l:Representation> </l:Variable> </l:VariableScheme> </l:LogicalProduct>
<pi:PhysicalInstance> <r:MaintainableID> <r:ID> 4245_PhysicalInstance </r:ID> </r:MaintainableID>
<pi:PhysicalDataProductReference> <r:Reference> <r:ID> 4245_PhysicalDataProduct </r:ID> </r:Reference> </pi:PhysicalDataProductReference>
<pi>DataFileIdentification> <r:IdentifiableID> <r:ID> 4245_DataFile </r:ID> </r:IdentifiableID>
<pi:Location> ICPSR </pi:Location>
<pi:URI> http://www.icpsr.umich.edu/cgi-bin/bob/newark?study=4245 </pi:URI> </pi>DataFileIdentification>
<pi:GrossFileStructure> <r:VersionableID> <r:ID> 4245_GrossFileStructure </r:ID> </r:VersionableID>
<pi:CaseQuantity> 1212 </pi:CaseQuantity>
<pi:OverallRecordCount> 1212 </pi:OverallRecordCount> </pi:GrossFileStructure>
<pi:Statistics> <pi:VariableStatistics> <pi:VariableReference> <r:Reference> <r:ID> V043015 </r:ID> </r:Reference> </pi:VariableReference>
<pi>TotalResponses> 942 </pi>TotalResponses>
<pi:CategoryStatistics> <pi:CategoryValue> </pi:CategoryValue>
<pi:CategoryStatistic> <pi:CategoryStatisticType> Frequency </pi:CategoryStatisticType>
<pi:Weighted> false </pi:Weighted>
<pi:Value> 270 </pi:Value> </pi:CategoryStatistic> </pi:CategoryStatistics>

```

Concept Universe

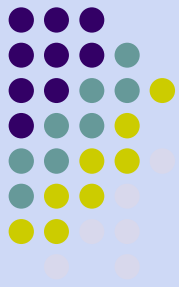
Question

Values
Value Labels
Variable name
Variable label

Statistics

Location:
Physical Data Product

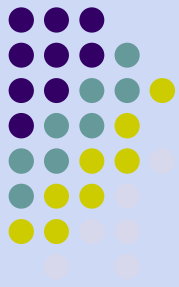
Where to start understanding the DDI?



DDI website: <http://www.ddialliance.org/ddi3/>

- **Sample XML Files:** created for the public review version of DDI 3.0
- **Getting Started with DDI 3.0:** describes how to begin coding documents in XML and how to convert existing collections.
- ***DDI Directions*:** newsletter on the DDI Alliance website

Help understanding and producing DDI 3.0 markup



- **DDI HELP:**

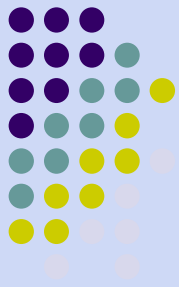
Detailed information about each module

<http://opendatafoundation.org/infocenter/index.jsp?topic=/net.sf.ddialliance.help/html/welcome.html>

- **Version 3.0 documentation:**

<http://www.ddialliance.org/ddi3/index.html>

Software to assist in document viewing, transformation and production



- **DeXtris:**

- XML browser to view and search DDI files
- Converts DDI 1/2 to DDI 3.0

<http://www.opendatafoundation.org/tools/dextris>

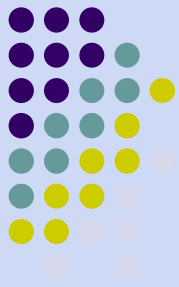
- **SPSS to DDI 3.0 converters**

- DDI Alliance website:

<http://www.ddialliance.org/DDI/ddi3/proof.html>

- StatsProgs2DDI (beta) from ZUMA:

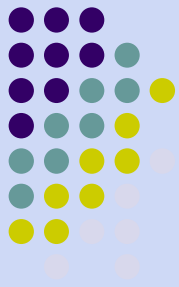
<http://db.zuma-mannheim.de/DDI/StatsProgs2DDI/StatsProgs2DDI.html>



Other XML tools

- **XML editor: oXygen**
 - Create new DDI instance
 - Edit/update DDI instance
 - Validate DDI instance
 - View schemas
- **DDI 3.0 Stylesheets/Transforms**
 - Basic: transform the XML file into XHTML for Web presentation.
 - Enhanced: passes the XML file through a series of stylesheets to add more advanced features to the XHTML display, such as graphical representation of frequencies, and automated calculation of valid percentages.

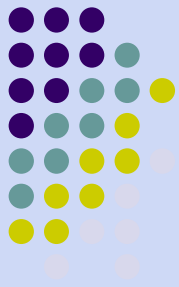
DDI Toolkit in development (open source)



Developed by the DDI Alliance, GESIS-ZUMA, UKDA, DDA, Canada RDCs, and OdaF.

- DDI Version 1/2 <-> Version 3.0 converters
- DDI 3.0 URN resolution tool
- DDI 3.0 validation tool
- Version 3.0 stylesheets with display and editing layers
- Grouping tool
- Concept management tool
- Registry applications

Tools from the International Household Survey Network



<http://www.surveynetwork.org/home/>

Combination of open source and NESSTAR (licensed)

Microdata Management Toolkit: uses the DDI metadata standard and the Nesstar content management and analysis system.

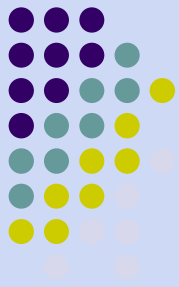
Metadata Editor is used to document data in accordance with international metadata standards (DDI and Dublin Core).

Explorer is a free reader for files generated by the Metadata Editor. It allows users to view the metadata and to export the data into various common formats (Stata, SPSS, etc).

CD-ROM Builder is used to generate user-friendly outputs (CD-ROM, website) for dissemination and archiving."

Quick Reference Guide for Data Archivists

DDI and Institutional Repositories



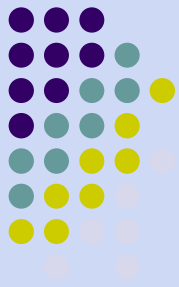
GOAL: Take advantage of DDI to classify, describe, and organize datasets

QUESTIONS:

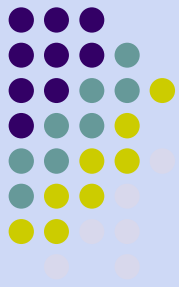
- use DDI 2 or DDI 3?
- how many and which tags are needed?
- does DDI address the requirements for access management, linking to other publications and preservation?
- in which ways could DDI be progressively incorporated and used in IRs?

DDI 1/2 or DDI 3.0?

Take what you can get!



- DDI 3.0 will not supersede DDI 2.1
- Both versions will
 - coexist
 - continue to be maintained
 - be used according to specific needs
- All DDI 1/2 markup will not have to be migrated to Version 3.0



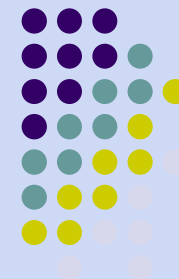
DDI 2 Lite

CESSDA 2001 recommendations:

<http://www.ddialliance.org/DDI/related/cessda-rec.pdf>

- Examples from the various CESSDA archives were collected and compared
- Defined a realistic least common denominator for the CESSDA archives in English
- Used to make cross archive substantive searches
- A set of strongly recommended fields constituting the basic information on any dataset.

DDI 2 and Dublin Core

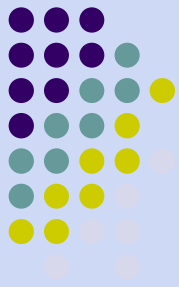


DC Element	DDI Element	Notes
Title	<titl> 2.1.1.1	Title of Data Collection
Creator	<AuthEnty> 2.1.2.1	Authoring Entity of Data Collection
Subject	<keyword> 2.2.1.1	Keyword(s)
	<topcClas> 2.2.1.2	Topic Classification
Description	<abstract> 2.2.2	Abstract
Publisher	<producer> 2.1.3.1	Producer of Data Collection
Contributor	<othId> 2.1.2.2	Other Identification/Acknowledgements - Data Collection
Date	<prodDate> 2.1.3.3	Production Date - Data Collection
Type	<dataKind> 2.2.3.10	Kind of Data
Format	<fileType> 3.1.5	Type of File
Identifier	<IDNo> 2.1.1.5	ID Number - Data Collection
	<holdings location="" callno="" URI=""> 2.1.8	Holdings Information - Data Collection
Source	<sources> 2.3.1.8	Sources - Used for Data Collection
Language		
Relation	<othrStdyMat> 2.5	Other Study Description Materials
Coverage	<timePrd> 2.2.3.1	Time Period Covered
	<collDate> 2.2.3.2	Date(s) of Data Collection
	<nation> 2.2.3.3	Country
	<geogCover> 2.2.3.4	Geographic Coverage
Rights	<copyright> 2.1.3.2	Copyright - Data Collection

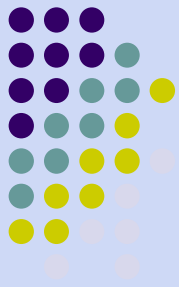
Dublin Core (Basic)	DDI 3.0
Contributor	Citation: Contributor (in Instance, StudyUnit, Group and Phys. Instance)
Coverage	Coverage: SpatialCoverage, and Coverage: TemporalCoverage (in Instance, StudyUnit, Group and Phys. Instance)
Creator	Citation: Creator (in Instance, StudyUnit, Group and Phys. Instance)
Date	Coverage: Temporal: AdministrativeDate, (in Instance, StudyUnit, Group and Phys. Instance) OR LifeCycleInformation: LifeCycleEvent: Date (in Archive)
Description	Abstract (in StudyUnit, and Group)
Format	Item: Format (in Archive)
Identifier	Item: Call Number (in Archive)
Language	Translation Information: Language (in Instance)

Dublin Core (Basic)	DDI 3.0
Publisher	Citation: Publisher (in Instance, StudyUnit, Group and Phys. Instance)
Relation	OtherMaterial (in Instance, StudyUnit, Group and Phys. Instance)
Rights	Citation: Copyright (in Instance, StudyUnit, Group and Phys. Instance)
Source	OtherMaterial (in Instance, StudyUnit, Group and Phys. Instance)
Subject	Coverage: Topical (in Instance, StudyUnit, Group and Phys. Instance)
Title	Citation: Title (in Instance, StudyUnit, Group and Phys. Instance)
Type	DataKind (in StudyUnit module)

Qualified Dublin Core elements and DDI 3



“Qualified” Dublin Core Elements	DDI 3.0
Audience	NA
Provenance	Item: OriginalArchiveOrganizationReference (in Archive)
RightsHolder	Citation: Copyright (in Instance, StudyUnit, Group and Phys. Instance) OR Organization/Individual: Role (In Archive: Organizations)

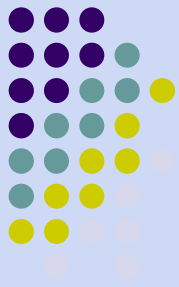


Dublin Core Elements	DC Element Qualifiers	DDI 3.0
Title	Alternative	Citation: Alternate Title
Description	Table of Contents	Item->Item (nestable) (in Archive)
	Abstract	Abstract
Date	Created Valid Available Issued Modified	Coverage: Temporal Coverage: Administrative Date: Type (has Controlled Vocabulary)
Format	Extent	Item: DataFileQuantity (in Archive)
	Medium	Item: Media (in Archive)
Relation	Is Version Of Has Version Is Replaced By Replaces Is Required By Requires Is Part Of Has Part Is Referenced By References Is Format Of Has Format	OtherMaterial: Type
Coverage	Spatial	Coverage: SpatialCoverage
	Temporal	Coverage: TemporalCoverage

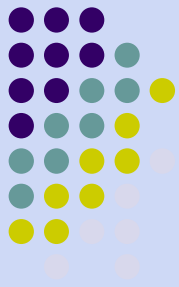
Source: Sanda Ianesu, ICPSR.

Core v3: based upon DDI 2 Lite

(see appendix in DataShare report)



- Describes the study as a whole, the data files, and the variables in the files
- Uses information from the following modules:
 - Study unit
 - Conceptual Component
 - Data Collection
 - Logical Product
 - Physical Data Product
 - Physical Instance
 - Archive
 - Organization



DDI 2 and OAI-PMH: ESDS

The Economic and Social Data Service (ESDS) has an OAI-PMH implementation that uses DDI 2

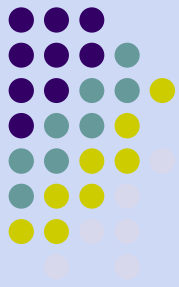
Responds to a *ListRecords* request returning a record with most of the information but not Data Files and Variables descriptions.

As an example, the (non public) URLs below display the metadata formats available in ESDS and the DDI record for study number 2473 :

<http://oai.esds.ac.uk/oai.asp?verb=ListMetadataFormats>

<http://oai.esds.ac.uk/oai.asp?verb=GetRecord&identifier=oai%3Aesds.ac.uk%3AESDS%2FESDSA%2Fsn2473.xml&metadataPrefix=ddi>

<http://oai.esds.ac.uk/>

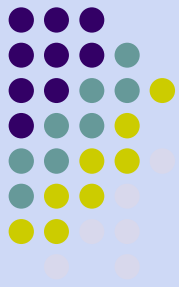


DDI 3 and OAI-PMH: MIT

Mackenzie Smith email, July 17, 2007:

- OAI-PMH Harvester Agent for Social Science Data Services
- This is a command-line executable agent that interacts with OAI Gateways or URLs to generate METS packages containing the resources referenced by the document residing at that URL.
- The object residing at that URL is a DDI document. In the DSpace case, a handler processes the DDI document instance found within an OAI /GetRecord/ call

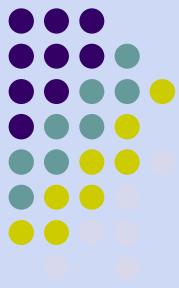
Using DDI for access management



- Both DDI 2 and DDI 3 have tags that can be used to deal with the access management requirements.
- In DDI 2 this can be achieved through the Data Access tag (`<dataAccs>`) which describes access conditions. In cases where access conditions differ across individual files or variables, multiple access conditions can be specified.
- In DDI 3.0 the relevant tags are in the Archive module under access type.

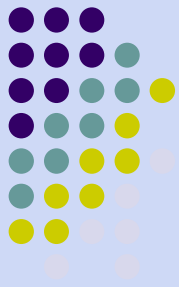
Using DDI for

Linking to other materials (publications, maps, images, etc.)



- Both versions of DDI will allow linking to publications.
- In DDI 2 the “Other Study Materials” section allows linking to related studies through title of study, author, producer, version and physical location amongst others.
- In DDI 3, linking to other materials happens in the study unit module.

Preservation metadata and DDI 3



PLEDGE:

Incorporate the Virtual Data Center into
PLEDGE activities

Transforming descriptive and administrative
metadata out of DDI formats and into MODS
and PREMIS for inclusion in METS SIPs.
<http://pledge.mit.edu/index.php/VDCIntegration>

The DDI instance can be imported into online data search and analysis systems such as NESSTAR, SDA or DataVerse

DDI metadata is created and enhanced for repository use; compliant metadata records are automatically produced from DDI XML files for other catalogues and repository harvesting sites such as Google Scholar or OAI-PMH; the DDI can be queried at the variable level (DDI 2) and across modules (DDI 3)

DDI XML file is stored in the repository and disseminated with the data; the IR provides a limited metadata record for the study

Researchers provide text files with information about the data; no DDI XML file is submitted to or disseminated from the repository; the IR provides a limited metadata record for the study.

